**Logo

Description automatically generated**

**OOTDiffusion: Outfitting Fusion based for Controllable Virtual Try-on**

by

**Phạm Anh Huy**

**Trịnh Như Phương**

**Phạm Quốc Vinh**

**THE FPT UNIVERSITY HO CHI MINH CITY**

**THE FPT UNIVERSITY HO CHI MINH CITY**

**AUTHOR CONTRIBUTIONS**

In the completion of this Final Capstone Project, each student has made significant and specific contributions to various aspects of the research. Their contributions are described as follows:

* **Conceptualization**, carried out by Phạm Anh Huy, laid the groundwork for the research idea and the overall development of the project.
* **Methodology**, developed by Phạm Anh Huy, shaped the approach for conducting the research.
* **Software**, with contributions from Nguyễn Phạm Quốc Vinh, involved developing and customizing the necessary software tools for the research.
* **Validation**, performed by Nguyễn Phạm Quốc Vinh and Trịnh Như Phương, ensured the accuracy of the data and methods.
* **Formal analysis**, conducted by Nguyễn Phạm Quốc Vinh, involved rigorous data analysis.
* **Investigation**, led by Trịnh Như Phương, guided the research process and data collection.
* **Resources**, managed and provided by Nguyễn Quốc Trung, included sourcing and utilizing necessary resources.
* **Data curation**, taken care of by Nguyễn Phạm Quốc Vinh, ensured the accuracy and organization of the collected data.
* **Writing—original draft preparation**, was the responsibility of Nguyễn Phạm Quốc Vinh, contributing to drafting the initial report.
* **Writing—review and editing**, also undertaken by Trịnh Như Phương, involved reviewing and refining the draft to ensure final quality.
* **Visualization**, created by Phạm Anh Huy, helped in visualizing the data and research findings.
* **Supervision**, by Trịnh Như Phương, provided guidance and oversight of the project's progress.
* **Project administration**, managed by Nguyễn Quốc Trung, ensured the smooth and efficient progression of the project.
* **Funding acquisition**, an essential task carried out by Trịnh Như Phương, involved seeking and managing finances for the project.

All authors have read and agreed to the final document of the Final Capstone Project. This is a testament to the close collaboration and commitment of each member towards the project's success.

**ABSTRACT**

Image-based virtual try-on (VTON) endeavors to produce a realistic depiction of a target individual adorned in a specific garment, constituting a formidable image synthesis endeavor that necessitates both the faithful representation of the clothed human form and the meticulous preservation of garment intricacies. To address this challenge, this study introduces Outfitting over Try-on Diffusion (OOTDiffusion), capitalizing on pretrained latent diffusion models and devising a novel network architecture tailored for authentic and adjustable virtual try-on simulations. Eschewing explicit warping procedures, our approach employs an outfitting UNet to glean garment detail features and seamlessly integrate them with the target human physique via a novel outfitting fusion mechanism within the denoising framework of diffusion models. Augmenting the versatility of our outfitting UNet, we introduce outfitting dropout during the training phase, affording the capacity to modulate garment feature intensity sans classifier guidance. Through extensive experimentation on the VITON-HD and Dress Code datasets, our method demonstrates its efficacy in generating high-fidelity outfitted images across diverse human and garment inputs, surpassing existing VTON methodologies in fidelity and adjustability, thus representing a notable advancement in virtual try-on technology.

**CONTENTS**

**ACKNOWLEDGMENTS** 1

**AUTHOR CONTRIBUTIONS** 2

**ABSTRACT** 3

C**ONTENTS** 4

**List of Figures** 5

**List of Tables** 6

**1. INTRODUCTION** 7

**2. RELATED WORK** 8

**3. PROJECT MANAGEMENT PLAN** 10

**4. MATERIALS AND METHODS** 11

**5.** **RESULTS** 16

**6. DISCUSSIONS** 17

**7. CONCLUSIONS** 19

**8.** **REFERENCES** 20

**9. APPENDIX**  21

List of Figures

(This section is used to list the name of the figures of the thessis as described below)

4.1Figure 1**.** This is a figure to describe datasets, with various input human and garment images 11

4.2 Figure 2.This is a figure to illustrates the pipeline of our OOTDiffusion 12

4.2. Figure 3. This is a figure to shows the attention maps with respect to the human body (1st row) and garment features (2nd row) aligned by our outfitting fusion 15

5.2. Figure 4. This is a figure to show qualitative results between the models trained on VITON-HD dataset and test on Dress Code dataset 17

5.2. Figure 5. This a figure to show result about Vietnamese actor Viet Anh with original image (left) and the fitted image (right) 17

List of Tables

(This section is used to list the name of the Table of the thesis as described below)

**3. Table 3.** Project plan 13

3. **Table 4.** Source codes and data 14

**1. INTRODUCTION**

In recent years, the intersection of deep learning and fashion technology has yielded innovative solutions that are transforming the way we engage with clothing in the digital age. One of the pioneering advancements in this domain is that Image-based virtual try-on (VTON) technology is gaining traction in the e-commerce sector, offering significant enhancements to consumer shopping experiences while lowering advertising expenses for apparel retailers. The objective of VTON is to create images of individuals wearing specific items of clothing, a goal that has spurred extensive research efforts toward achieving more realistic and accurate renditions.

Presently, VTON technologies confront two primary hurdles. The first challenge involves producing images that are sufficiently lifelike and natural to prevent any discordance, with recent studies often employing generative adversarial networks (GANs) or latent diffusion models (LDMs) for this purpose. Although GAN-based approaches have struggled with accurately rendering garment folds and realistic body portrayals, LDM-based techniques have shown promise in enhancing the authenticity of the virtual try-on images. The second major issue is retaining intricate garment details such as text, textures, colors, patterns, and lines during the try-on simulation. Traditional methods typically involve a warping process to adjust the garment's appearance to fit the model's body shape, a step heavily reliant on the precision of the warping technique and prone to overfitting. Conversely, some newer LDM approaches attempt to integrate garment details through CLIP textual inversion, yet often fail to maintain the nuanced aspects of clothing.

In response to these challenges, our research introduces a novel LDM-based VTON method, dubbed Outfitting over Try-on Diffusion (OOTDiffusion), which utilizes pre-trained latent diffusion models for generating highly realistic images and incorporates an innovative outfitting UNet to capture garment details in one step. Furthermore, our method includes an outfitting fusion process for accurately merging garment features with various human forms and poses, alongside an outfitting dropout technique to enhance the method's adaptability to garment features. This comprehensive approach not only addresses the existing challenges in VTON but also establishes new standards for realism and control in virtual try-on solutions, as demonstrated by our extensive testing on high-resolution benchmark datasets, where our method outperforms current leading VTON technologies.

**2. RELATED WORK**

For several years, the field of image-based virtual try-on (VTON) has aptivated researchers due to its potential to revolutionize online shopping and the fashion industry. This domain has seen a variety of approaches aimed at enhancing the realism and accuracy of virtually trying on garments. Notably, advancements have primarily hinged on the use of generative adversarial networks (GANs) and latent diffusion models (LDMs) to craft convincing images of individuals donned in selected attire and impressive results in creating photo-realistic images

In the sphere of GAN-based VTON solutions, initiatives like VITON-HD have been pioneering, introducing high-resolution datasets alongside novel methodologies such as ALIAS normalization and a specialized generator to tackle issues of misalignment between the digital garments and the intended areas on the target figures. The area of human-centric image synthesis has seen considerable advancements with the application of Generative Adversarial Networks (GANs), particularly models based on StyleGAN, which have produced impressive results in creating photo-realistic images. Within this realm, specific efforts like InsetGAN and StyleGAN-Human have been directed towards synthesizing high-fidelity human figures by integrating outputs from various pre-trained GANs, each responsible for different body parts, or by optimizing factors such as dataset size, distribution, and alignment. Other contributions include HR-VITON, which innovatively combined warping and segmentation processes to mitigate challenges associated with body occlusions and garment alignment. Similarly, GP-VTON introduced a novel warping module and a distinctive training strategy aimed at enhancing the deformation of garments for a more natural fit.

Despite these advancements, GAN-based approaches often encounter limitations related to the warping process, particularly in maintaining the integrity of garment folds, lighting, and shadow effects. This can significantly impact the visual fidelity of the try-on results. Furthermore, these methods tend to overfit training datasets, resulting in diminished performance when applied to images outside of the training distribution.

LDM-based methods have also been explored, with initiatives like LaDI-VTON and DCI-VTON employing warping processes. These methods have sought to preserve garment details by mapping visual features to token embedding spaces or by integrating warped garments with person images for subsequent refinement through diffusion models. However, challenges persist in fully retaining the intricate details of garments, with some approaches suffering from information loss attributed to encoding processes.

More recent approaches, such as StableVITON, have attempted to circumvent these issues by eliminating the need for independent warping, instead relying on novel architectural elements to understand the semantic relationships between garments and human figures. Yet, these solutions also face obstacles, including increased computational demands and potential information loss within their models.

Contrasting with previous methodologies, our proposed system, OOTDiffusion, leverages LDMs without the need for preliminary warping steps. By fine-tuning a pre-existing outfitting UNet, we achieve a seamless integration of garment details into the virtual try-on process. Our approach, outfitting fusion, effectively merges these details into the denoising UNet with minimal loss of information, showcasing a significant improvement in the preservation of garment fidelity.

In addition to exploring VTON, recent research has delved into enhancing the controllability of LDMs in various image-generation tasks. Techniques such as Prompt-to-Prompt and Null-text Inversion have demonstrated the capacity to fine-tune image outputs by manipulating input captions, thereby bypassing the need for additional model training. Other noteworthy developments include InstructPix2Pix, which generates edited images based on text instructions, and Paint-by-Example, which adopts a self-supervised approach for granular image manipulation.

Our work aligns with these endeavors by emphasizing the controllability aspect in the VTON domain. Through outfitting fusion and dropout techniques, we enable precise adjustments to the garment features within the generated images, thus offering a novel and effective solution to the challenges of image-based virtual try-on. This advancement not only contributes to the academic discourse but also promises substantial implications for the future of e-commerce and online fashion retailing.

**3. PROJECT MANAGEMENT PLAN**

[Provide the overall project objective description and then the specific target metrics of your project in term of quality, time, and cost (allocated effort distribution for project activities]

**Table 3.** Project plan

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Task name** | **Priority** | **Owner** | **Start date** | **End date** | **Status** | **Issues** |
| Find documents | High | Phạm Anh Huy | 05/01/2024 | 10/01/2024 | completed | None |
| Review papers | High | Trịnh Như Phương | 08/01/2024 | 11/01/2024 | completed | None |
| Review and analyze public dataset | High | Nguyễn Phạm Quốc Vinh | 11/01/2024 | 20/01/2024 | completed | datasets do not match |
| Collect and label data | Medium | Nguyễn Phạm Quốc Vinh | 20/01/2024 | 30/01/2024 | completed | None |
| Experiment | High | Phạm Anh Huy | 01/02/2024 | 15/02/2024 | completed | n other languages, many functions are too old |
| Compare results | Medium | Trịnh Như Phương | 16/02/2024 | 29/02/2024 | completed | First, since our models are trained on paired human and garment images, it may fail to get perfect results for cross-category virtual try-on, e.g., to put a T-shirt on a woman in a long dress, or to let a man in pants wear a skirt. Another limitation is that some details in the original human image might be altered after virtual try-on, such as muscles, watches or tattoos, etc. |
| Writing appendix | Medium | Nguyễn Phạm Quốc Vinh | 01/03/2024 | 04/03/2024 | completed | None |
| Future work | High | Phạm Anh Huy | 01/04/2024 | … | Planned | None |

The supporting information can be downloaded at: www.abc.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title. By the following table

**Table 4.** Source code and data

|  |  |  |
| --- | --- | --- |
| **Items** | **Link** | **Description** |
| Data | Link | [VITON-HD](https://github.com/shadow2496/VITON-HD)  [Dress Code](https://github.com/aimagelab/dress-code) |
| Source code | Link | [OOTDiffusion](https://github.com/ChrisPham-0502/OOTDiffusion-for-Virtual-Try-on.git) |

**4. MATERIALS AND METHODS**

**4.1 Materials**

****

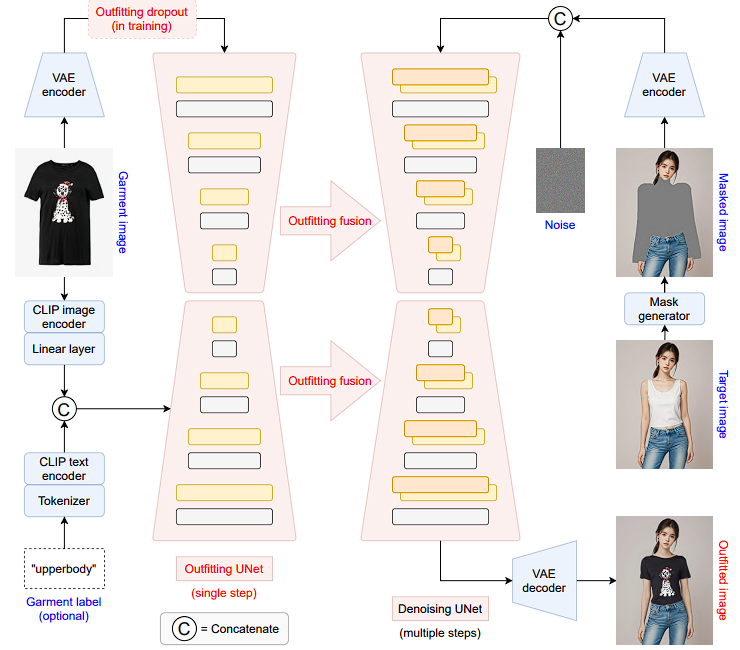
**Figure 1.** Outfitted images (1024 × 768) generated by our OOTDiffusion trained on the VITON-HD [6] (1st row; supporting upper-body garments) and Dress Code [33] (2nd row; supporting upper-body garments, lower-body garments and dresses) datasets, with various input human and garment images.

**Datasets:** We utilize two popular datasets in the Virtual Try-On problem for training our OOTDiffusion model: VITON-HD and Dress Code. We prioritize using the resolution of 1024\*768 to enable the model to effectively learn in handling image generation and reduction of noise – which will be elaborated on later in the model section.

**Stable Diffusion:** Our OOTDiffusion model is built upon the architecture of Stable Diffusion, one of the commonly used diffusion models in image generation tasks. Stable diffusion employs an autoencoder (VAE) consisting of an encoder Ɛ and a decoder Ɗ to process and train images in a latent space. Within this latent space, the model utilizes a trained UNet to denoise images – noise generated from the Gaussian Noise algorithm – with the aim of returning semantically meaningful images as input for the VAE decoder.

**CLIP:** In addition to training the model with image data, OOTDiffusion also utilizes CLIP – a text encoder – to encode input prompts. The model relies on these prompts to classify clothing components into upper body, lower body, and dress segments to better fit clothes onto the wearer accurately.

**4.2 Methods**



**Figure 2.** Overview of OOTDiffusion pipeline. The garment image is encoded and processed through outfitting UNet, incorporating features from CLIP encoders. Garment latents undergo outfitting dropout during training for classifier-free guidance. The human image is masked and combined with Gaussian noise, then passed through denoising UNet for multiple sampling steps. The resulting feature map is decoded into the try-on result.

**Overview:** Fig. 2 illustrates the pipeline of our OOTDiffusion. Given a target human image x ∈

ℝ3\*H\*W and an input garment image g ∈ ℝ3\*H\*W, OOTDiffusion model is able to generate a realistic outfitted image xg g ∈ ℝ3\*H\*W. In the architecture, we employ OpenPose [4,5,46,51] and HumanParsing [28] to generate a masked human image xm ∈ ℝ3\*H\*W , then use a VAE encoder Ɛ to transform it into the latent space as Ɛ(xm) ∈ ℝ4\*h\*w , where and . Then we concatenate Ɛ(xm) with a Gaussian noise ε ∈ ℝ4\*h\*w as the input zT ∈ ℝ8\*h\*w for denoising UNet. Note that we add 4 zero-initialized channels to the first convolutional layer of the denoising UNet to support our input with 8 channels. On the other side, we feed the encoded garment latent Ɛ(g) ∈ R4\*h\*w into an (i) outfitting UNet to learn the garment features in a single step, and integrate them into the denoising UNet via our (ii) outfitting fusion. And we perform (iii) outfitting dropout for Ɛ(g) particularly in the training process. In addition, we also conduct CLIP textual-inversion [10] for the garment image g, and optionally concatenate it with a text embedding of the garment label y ∈ {“upperbody”, “lowerbody”, “dress”} as an auxiliary conditioning input, which is fed into both outfitting and denoising UNets via the cross-attention mechanism [48]. Finally, after multiple steps of the denoising process, we use a VAE decoder Ɗ to transform the denoised latent z0 ∈ ℝ4\*h\*w back into the image space as the output image xg = Ɗ(z0) ∈ ℝ3\*H\*W. We will elaborate the key technologies (i.e., (i) outfitting UNet, (ii) outfitting fusion, and (iii) outfitting dropout) of our OOTDiffusion in the following sections.

**Outfitting UNet:** As introduced above, we propose an outfitting UNet to efficiently learn the detail features of the garment image g. The left side of Fig. 2 shows the architecture of our outfitting UNet, which is essentially identical to the denoising UNet of Stable Diffusion. The encoded garment latent Ɛ(g) ∈ ℝ4\*h\*w is fed into the outfitting UNet ωθ′, and then incoporated into the denoising UNet ϵθ via our outfitting fusion (see the next section). Along with the aforementioned auxiliary conditioning input, the outfitting and denoising UNets are jointly trained by minimizing the following loss function:

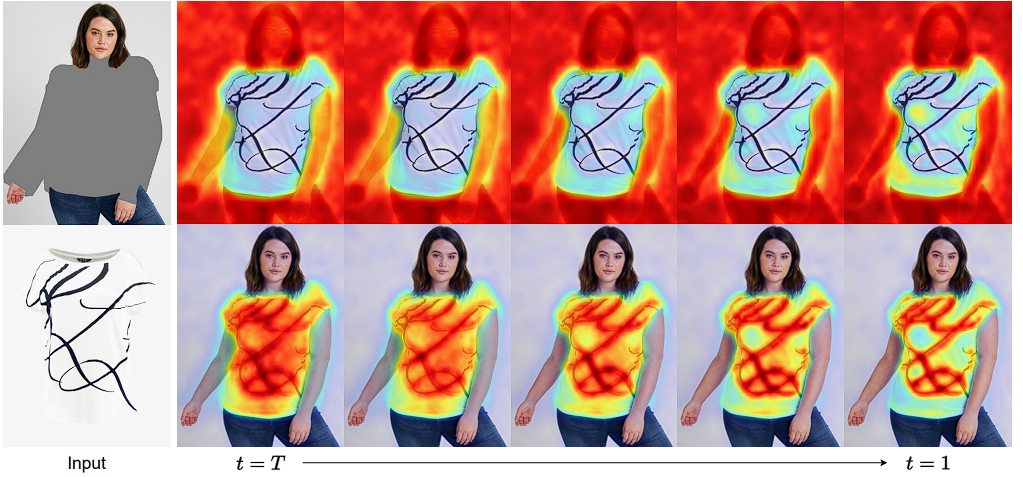


where ψ = τg(g) © τy(y) represents the auxiliary conditioning input for both ωθ′ and ϵθ. While τg and τy refer to the pretrained CLIP image encoder and text encoder respectively, and ○c denotes concatenation. In practice, we directly duplicate the pretrained UNet weights of Stable Diffusion [40] for the initialization of both our outfitting and denoising UNets (except for the zero-initialized channels added to the first convolutional layer), and jointly finetune them on the high-resolution VTON datasets [6, 33]. Note that ωθ′ and ϵθ do not share any weights in the training process. We claim that our tactical utilization of the pretrained models dramatically improves the training efficiency and reduces the training cost. Moreover, compared with the denoising UNet, a significant difference in our outfitting UNet is that it requires only one step forward process before the multiple denoising steps in inference, causing a minimal amount of extra computational cost to the original Stable Diffusion [40].

**Outfitting Fusion:** Based on our proposed outfitting UNet and inspired by the spatial-attention mechanism [21,48], we propose an outfitting fusion process to incorporate the learned garment features into the denoising UNet. First, we dive into the transformer blocks [48] of two UNets, finding each pair of feature maps used as the input to the corresponding self-attention layers [48]. Given the nth pair of the feature maps gn, xn ∈ ℝcn\*hn\*wn, we concatenate them in the spatial domain as:



And we replace xn with the concatenated feature map xgn as the input to the self-attention layer of the denoising UNet. Then we crop out the fist half of the output feature map as the final output of the self-attention layer. Fig. 3 visualizes the attention maps learned in our modified self-attention. We observe that the unmasked region focuses attention on the human body (1st row), and the masked pixels are attracted to the garment features (2nd row). Meanwhile, during the denoising process, the attention to the human body gradually includes part of the masked region like the neck and arms, and the attention to the garment features gradually increases in the region of the complicated patterns. Through outfitting fusion in the self-attention layers, the garment features are implicitly warped and effectively correlated with the target human body with negligible information loss. Hence the denoising UNet is made capable of learning the precise features from the outfitting UNet for preserving garment details and naturally adapting them to the target human body in the generated image.



**Figure 3.** Visualization of Outfitting Fusion shows the attention maps with respect to the human body (1st row) and garment features (2nd row) aligned by our outfitting fusion.

**Outfitting Dropout:** In order to further enhance the controllability of our VTON method, we employ an outfitting dropout operation in training to enable classifier-free guidance [20] with respect to the garment features. Classifier-free guidance has been broadly used in conditional image generation [3, 35, 44, 56] for trading off the quality and diversity of images generated by latent diffusion models. Specifically in the training process of our outfitting UNet, we randomly drop the input garment latent as Ɛ(g) = Ø, where Ø ∈ ℝ4\*h\*w refers to an all-zero latent. In this way, the denoising UNet is trained both conditionally and unconditionally, i.e., with and without the outfitting fusion. Then at inference time, we simply use a guidance scale s**g** ≥ 1 to adjust the strength of conditional control over the predicted noise as:



where we omit some minor terms compared with Eq. (2) for the sake of brevity. In practice, we empirically set the outfitting dropout ratio to 10% in training, i.e., 10% of garment latents Ɛ(g) are set to Ø. And the optimal value of the guidance scale sg is usually around 1.5 ∼ 2.0 according to our ablation study (see Sec. 4.3). Fig. 4 and Tab. 1 demonstrate the effects of our outfitting dropout and different guidance scale values.

**5. RESULTS**

*5.1. Experiments*

* **Datasets.** As mentioned above, our experiments are performed on two high-resolution (1024 × 768) virtual try-on datasets, i.e., VITON-HD [6] and Dress Code [33]. The VITONHD dataset consists of 13,679 image pairs of frontal half-body models and corresponding upper-body garments, where 2032 pairs are used as the test set. The Dress Code dataset consists of 15,363/8,951/2,947 image pairs of full-body models and corresponding upper-body garments/lower-body garments/dresses, where 1,800 pairs for each garment category are used as the test set.
* **Evaluation Metrics.** We evaluate the results in both the paired and unpaired settings, where the paired setting provides the target human and the corresponding garment images for reconstruction, and the unpaired setting provides the different garment images for virtual try-on. Specifically for Dress Code [33], we note that the evaluation is performed on the entire dataset rather than being limited to upper-body garments. This more effectively validates the feasibility of each method in real-world applications with various garment types. In the quantitative evaluation, though our OOTDiffusion supports higherresolution (1024 × 768) virtual try-on, all the experiments are conducted at the resolution of 512 × 384 for fair comparison with previous VTON methods. For the paired setting, we use LPIPS [58] and SSIM [50] to measure the quality of the generated image in terms of restoring the original image. For the unpaired setting, we employ FID [19] and KID [2] for realism and fidelity assessment. We follow the previous work [7, 32, 37] to implement all of these metrics.

*5.2 Evaluations*

In our experiments, we initialize the OOTDiffusion models by inheriting the pretrained weights of Stable Diffusion v1.5 [40]. In ordinary paper, they show the compared results between models trained on VITON-HD dataset with OOTDiffusion, we can see the qualitative result that our model is more optimal and precise. The Fig. 4 will show below.



**Figure 4.** Qualitative results between the models trained on VITON-HD dataset and test on Dress Code dataset.

We also tested the model on Vietnamese artist images to ensure various inputs and outputs of OOTDiffusion. The visualization is shown below.



**Figure 5.** The result about Vietnamese actor Viet Anh with original image (left) and the fitted image (right).

**Table 1.** The study of different guidance scale values on the VITON-HD dataset [6].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Guidance scale** | **LPIPS** | **SSIM** | **FID** | **KID** |
| 1.0  1.5  2.0  2.5  3.0  5.0 | 0.0749  0.0705  0.0708  0.0746  0.0753  0.0788 | 0.8705  0.8775  0.8766  0.8691  0.8684  0.8640 | 8.99  8.81  8.80  8.84  8.95  9.28 | 0.89  0.82  0.86  0.89  0.96  1.22 |

**6. DISCUSSION**

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

**7. CONCLUSIONS AND PERSPECTIVES**

This section is used for Summarize your report in one or two paragraphs. Which methods or algorithms achieves the best performance? Why do you think that some algorithms worked better than others? Is there any limitation and drawback of your works? If you had more time and more computational resources, or more team members, what aspects you should propose and explore to improve the performance?

**CONFLICTS OF INTEREST:** Declare conflicts of interest or state “The authors declare no conflict of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results”.

**8. REFERENCES**

1. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J.,  
Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf 2(3), 8 (2023)  
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)  
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)  
4. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)  
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)  
6. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14131–14140 (2021)  
7. Detlefsen, N.S., Borovec, J., Schock, J., Jha, A.H., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M., Falcon, W.: Torchmetrics-measuring reproducibility in pytorch. Journal of Open Source Software 7(70), 4101 (2022)  
8. Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., Yin, J.: Towards multi-pose guided virtual try-on network. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9026–9035 (2019)  
9. Fenocchi, E., Morelli, D., Cornia, M., Baraldi, L., Cesari, F., Cucchiara, R.: Dual-branch collaborative transformer for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2247– 2251 (2022)  
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)  
11. Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., Luo, P.: Disentangled cycle consistency for highly-realistic virtual try-on. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16928–16937 (2021)  
12. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8485–8493 (2021)  
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM 63(11), 139–144 (2020)  
14. Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7599–7607 (2023)  
15. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10471–10480 (2019)  
16. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7543–7552 (2018)  
17. He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3470–3479 (2022)  
18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)  
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017)  
20. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)  
21. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)  
22. Issenhuth, T., Mary, J., Calauzenes, C.: Do not mask what you do not need to mask: a parser-free virtual try-on. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 619–635. Springer (2020)  
23. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)  
24. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. arXiv preprint arXiv:2312.01725 (2023)  
25. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)  
26. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)  
27. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision. pp. 204–219. Springer (2022)  
28. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(6), 3260–3271 (2020)  
29. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)  
30. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P., Lai, Y.K.: Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In: CVPR Workshops. vol. 3, pp. 10–14 (2020)  
31. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023) OOTDiffusion for Controllable Virtual Try-on 17  
32. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladivton: Latent diffusion textual-inversion enhanced virtual try-on. arXiv preprint arXiv:2305.13501 (2023)  
33. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2231– 2235 (2022)  
34. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2iadapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)  
35. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)  
36. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)  
37. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11410–11420 (2022)  
38. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)  
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)  
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)  
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)  
42. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)  
43. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)  
44. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic textto-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems 35, 36479–36494 (2022)  
45. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image superresolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(4), 4713–4726 (2022)  
46. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)  
47. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)  
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)  
49. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristicpreserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV). pp. 589–604 (2018)  
50. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 3(4), 600–612 (2004)  
51. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)  
52. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23550–23559 (2023)  
53. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)  
54. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7850–7859 (2020)  
55. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)  
56. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 2(3), 5 (2022)